

A Logistic Regression Model of Irony Detection in Chinese Internet Texts

Frank Z. Xing and Yang Xu

Department of Information Management, Peking University, Beijing, China
{xzt, yang.xu}@pku.edu.cn

Abstract. The research of sentiment analysis has become fascinating with the support of emerging Internet language material. In this paper, irony in Chinese is investigated as a sentiment that has not been meticulously studied. We describe here a set of features and their computational formalization for detecting irony at a linguistic level. Comments from online forum are collected and detected whether ironical or not, with a logistic regression model. The efficacy and validity of our model is proved by comparison with other popular learning methods and statistical testing. The model achieves a performance close to state-of-the-art results in English and Italian from recall and accuracy perspective.

Keywords: sentiment analysis, formalized features, intentional meaning, logistic regression

1 Introduction

Irony is an ubiquitous phenomenon in many natural languages. There has been a long-time discussion in English rhetorics about irony as a figure of speech. In the discussion the classification of irony and the widely accepted definition of irony as “saying the opposite of what is meant” are proposed [1]. With the rapid development of Internet content, the task of irony detection gains significance in both sentiment analysis and its practice, such as opinion mining. However, the topic of irony has received little serious computational treatment in the previous discussion. Thus, works emphasized on sentiment analysis start to employ the method of summarising and elaborating linguistic features of irony for computational formalization [2–4].

However, most of previous studies are dominated by irony detection in English contexts, other languages have been neglected for a long time. Although [5, 6] starts to discuss automated irony detection in Italian and Brazilian Portuguese, there is still a lack of literature dealing with non Indo-European languages. While irony is a pervasive linguistic phenomenon, some of the specific features vary with cultural and structural properties of language. Transplanting all the features of irony in English to other languages as a whole will meet technical difficulties, and the performance of the approach will be weakened. To solve this problem, this paper, accordingly aims to explore characteristic features of irony in Chinese contexts, and improve the detection techniques.

The approach to detect irony in Chinese goes in three steps. First, theories on irony are systematically reviewed to clarify the main features of irony, and their applicability in Chinese are validated. Second, key features are summarized and utilized for computational formalisation to form independent variables. Third, a logistic regression model with threshold is employed to synthesize the features for a better-performanced solution of irony detection. Since the automatic detection of irony in Chinese has not been well studied, we take the state-of-the-art results in English, Portuguese and Italian for comparison. The approach provides new insights on computational formalisation of irony. Meanwhile, through the evaluation part, some susceptible hypotheses are tested, hence our knowledge on Chinese irony from a computational linguistic perspective is extended.

The rest of this paper is organized as follows. Section 2 introduces related works on irony theories and and illustrates typical examples of irony sentences in Chinese. Section 3 describes the data we extracted from Internet forums and the formalisation of features in detail. The performance of the model is discussed and evaluated in Section 4, highlighting its validity. Section 5 is the conclusion.

2 Irony in Chinese contexts

Theorist [7] suggests that identification of irony, even in face-to-face communication can be difficult due to perception of irony varying among peoples. Many people realize irony while others not, because of the ambiguity of language. The difficulties in irony perception of human indicate the complexity of computer based irony detection.

Irony can be understood from three perspectives, as a figure of speech, as mention [3] and as phrasal pattern [8]. Traditionally, irony as a figure of speech, has been mainly characterised as negation that conveys the original meanings of utterance. However, Grice in [9] finds out violations as another feature in irony. He defines irony as “a case of flouting the cooperative principle, by violating the maxim of quality”. Giora in [10] elaborates Grice’s violations into indirect negation and graded salience. The former one refers to the linkage between the literal and implied meanings are indirect but negative. However, the latter one argues that irony text can be understood polysemously whereas irony occurs when salience meanings have been processed by speakers over other (mainly literal) meanings. A further development of this theory is to understand irony as an indirect speech act [8]. Such theory regards irony as an intentional expression of insincerity that cannot be limited to assertions but can also be applied to many other occasions such as congratulations.

Another theory about irony concerns contextual-appropriateness, which points out that rather than violating maxim of quality, irony shall be defined as mention where utterance seems to be literally not appropriate to its context [3]. Except for unintentional irony, ironical utterance can be inappropriate but relevant to its contexts. The speaker intentionally knows such inappropriateness while he/she intends the audiences are also aware of these as well. Irony is not necessarily negative, but utterance with inappropriateness from this angle. This

theory extends the boundary of irony and suggests that understanding of irony requires two steps that includes (1) understanding the intentional meaning of irony, which is probably through graded salience and (2) assessing the semantic conflicts.

Partington in [11] also mentions irony patterns in phrase usage, mainly by combining elements within the phrase of opposing evaluative polarity. This is named phrasal irony. However, in Chinese the concept of “phrase” is more free than in English. Thus we make extend the range of polarity to a whole sentence.

To conclude, theories based on English contexts have provided a relatively comprehensive map of irony. Key descriptive features include (1) relevant but inappropriate to its contexts, (2) intentionally or unintentionally perceived inappropriateness and (3) with certain phrasal patterns. Reyes et al. in [12, 13] proposed corresponding formalizable dimensions called *Signatures*, *Unexpectedness* and *Emotional scenarios* and the sub-branches.

The dimensions are basically applicable to irony in Chinese. However, features can be different. For example, Reyes et al. in [13] suggested capitalized words as a hint of pointness, but this phenomenon does not exist for Chinese characters. Emoticons and punctuation marks are more appropriate for emotional scenarios. Thus in Chinese, *signatures* mainly consists of counter-factuality in [14]. Example 1 in Table 1¹ illustrates the “clever” does not carry its literal meaning, or he would not be uncomfortable with his father’s words and parental love. A general idea of irony in Chinese is that, irony is presented as text whose signifying is different from its signified. In other words, the literal meaning of the text is different from the real semantic meaning the speaker intended to express. Unexpectedness suggest that the collocation is not common. In Chinese the collocation can be sentence level inappropriate. For instance in Example 2, Chinese cruller can not be normally modified by adjectives like “slick” and “sly”.

Table 1. Typical Examples of Irony in Chinese

Example 1	我那时真是聪明过分，总觉他说话不大漂亮。 I was too clever at that time, as I never thought he spoke appropriately.
Example 2	我不欺不瞒地说，他是个油条，八面见光，是个玲珑灯。 I will not cheat you, he is a Chinese cruller, slick and sly.
Example 3	地沟油也是民生工程之一啊！ Gutter oil is also one of the service offered to the public!

Chinese scholars [15, 16] also point out that irony usually appears in contexts with certain emotion. For instance, Example 1 shows an emotion of regretting and self-accusation. Pragmatically speaking, the contraction in meanings is usu-

¹ Example 1 is from Chinese essayist Zhu Ziqing’s famous prose “My Father’s Back”; Example 2 is from Zhu Chunyu’s novel “In a Sea of People”; Example 3 is in our dataset extracted from the Internet.

ally used to express certain emotions, to help better highlight some traits of the objects. Therefore, tracing emotional fluctuation in certain contexts can be a good indicator of expression of irony. In Chinese, this emotional scenario is dominated by coexistence of commendatory and derogatory terms. There have been researches about polarity [17, 18]. Sentence level polarity well represent the emotional scenario.

3 Model

3.1 Dataset for Irony Detection

In consistent with previous research, the dataset used is extracted from the Internet. However, Twitter is not the most popular platform for disseminating and sharing information and opinions in Chinese. In order to make a better coverage and representativeness, the data comes from three popular Chinese news forums – Tencent News, Sina Social News and Tianya Forum.

From these news forums, we choose 15 topic articles from the most popular chart (ranked by page view) in September 2013 and September 2014, each of which is followed by thousands of feedback and comments. Because irony is not a frequent phenomenon in natural language, we filtered irrelevant comments to increase the odds of irony. Literature [19] suggests by incorrectly classifying all 2,795 ironic simile speeches as non-ironic, the system still achieves an overall F-measure of 0.81. Such a system has no irony detection capabilities at all, yet achieves reasonable performance just because of the imbalance of non-irony to irony among speeches.

As a result, 2,602 comments in total are extracted from bellowing these news articles. 416 comments are marked positive of irony according to the survey, which suggests an overall ratio of 16.0%. The character count for each comment ranges from 4 to 191, with a variance of 611.7. Therefore comment length is regularized in the following formalization.

3.2 Computational Formalization of Irony Features

A speech is defined as a stream of meaningful text that carries emotion and intention. The specific form of a speech in the experiment is a short online comment. The intentional meaning of a comment and the article it belongs to are related. Theme refers to the literal meaning of an article. The intentional meaning of a comment is assumed the same as the theme of its related article.

Take that U.S. president Barack Obama refuted claims that hostility to his presidency was due to racism, by noting that “I was actually black before the election” as an example [19]. This statement was both literally true and ironic. In this case the literal meaning can be concluded directly from his statement. However, if we do not know about the context, it would be extremely difficult to summarize the intentional meaning. In our experiment, we fortunately capture the intentional meaning by summarizing the article, which is highly probably

about racism, indicating irony by the difference between literal meaning and intentional meaning.

Inspired by automatic indexing, we utilized the keyword vector of a text to represent its literal meaning. Then the vector angle between the keyword vector of a speech and its theme is a good metric of the deviation from the speech's intentional meaning to its literal meaning.

We use T_i to denote a certain theme shared by several speeches $S_1^i, S_2^i, \dots, S_j^i$. i denotes the number of articles and j denotes the number of comments related to T_i .

$\alpha(T_i) = (w_1 K_1^{T_i}, w_2 K_2^{T_i}, \dots, w_m K_m^{T_i})$ denotes the weighted keyword vector of theme T_i . $\alpha(S_j^i) = (w_1 K_1^{S_j^i}, w_2 K_2^{S_j^i}, \dots, w_n K_n^{S_j^i})$ denotes the vector of speech S_j^i . K_i is the i^{th} keyword, w_i is the normalised weight of K_i .

Feature 1: $Re(T_i, S_j^i) = \cos \langle \alpha(T_i), \alpha(S_j^i) \rangle$ measures the similarity of literal meaning and intentional meaning of S_j^i . Therefore $Re(T_i, S_j^i)$ is negatively related to the probability that S_j^i expresses irony. In the calculation, we regard $|\alpha(T_i)|$ as a constant because the theme apparently generates more keywords to make $|\alpha(T_i)|$ far more large than $|\alpha(S_j^i)|$. Consequently we are only concerned by the length of a speech.

$$Re(T_i, S_j^i) = \frac{\alpha(T_i) \cdot \alpha(S_j^i)}{|\alpha(T_i)| |\alpha(S_j^i)|} = \frac{1}{|\alpha(S_j^i)|} \sum_{K_m^{T_i} = K_n^{S_j^i}} w_m^T w_n^S \quad (1)$$

Feature 2: If the word vector² of S_j^i has the form $(w_1^S, w_2^S, \dots, w_q^{S_j^i})$, $I(S)$ can be calculated as the number of inappropriate collocation $I(S) = \mathcal{N}(w_k^S, w_{k+a}^S), k \in \{1, 2, \dots, q - a\}$. The window length a can be adjusted to balance the probability of missing and mistaken inappropriate collocation.

Feature 3: If each word is assigned a value to denote its polarity, the corresponding emotion vector can be written as $\epsilon(S_j^i) = (E_1^{S_j^i}, E_2^{S_j^i}, \dots, E_q^{S_j^i})$. Let $E^+ = |\sum_{r=1}^q E_r^{S_j^i}|, \forall E_r^{S_j^i} > 0$; $E^- = |\sum_{r=1}^q E_r^{S_j^i}|, \forall E_r^{S_j^i} < 0$. If $E^+ E^- \neq 0$, $E(S_j^i) = \frac{E^+ + E^-}{1 + \rho|E^+ - E^-|}$, otherwise $E(S_j^i) = \sigma(E^+ + E^-)$, ρ and σ are coefficients for smoothing.

To achieve the computational formalisation aforementioned, several techniques are employed. First, the software package, MMSeg for Java lucene Chinese analyzer, is used to deal with Chinese word segmentation. The maximum

² note the difference between “word vector” and “keyword vector”, which is ignoring the weight values in $\alpha(S_j^i)$, A keyword vector of S_j^i is obtained as a vector $(w_1^S, w_2^S, \dots, w_p^S)$.

matching algorithm that implemented the conversion from S_j^i to $(w_1^S, w_2^S, \dots, w_p^S)$ is introduced in [20]. Second, we implemented the keyword vector generation algorithm introduced in [21]. The algorithm first builds a graph of the annotated sentences, then do recursions to estimate the weight of each edge. At the final stage weights are assigned to the words and those words whose weight equals to zero are deleted. Third, we referred to the method of fundamental sentiment word polarity computation in [18] to modify the emotion dictionary we constructed, in which each word is quantitatively described with its emotional intention, part of speech and its synonyms³. The calculation of $E(S_j^i)$ is based on the modified dictionary.

Here is an example of how the formalisation of features works. In our dataset, there is a theme article titled “星巴克：在中国更贵因为顾客长时间滞留店内喝咖啡 (Starbucks: More expensive in China for clients stay several hours to drink coffee)”. The TextRank algorithm generates $\alpha(T)$ as (1* China, 0.44* Starbucks, 0.41* America, 0.25* coffee, 0.23* café, ..., 0.1* (TV station), 0.1* quite, 0.1* claim, 0.1* (raw material)). A speech of this theme, apparently annoyed by the story of Starbucks, is “Starbucks is just a symbol of American junk food culture”. $\alpha(S)$ is calculated as (1* symbol, 1* Starbucks, 0.1* junk, 0.1* culture). Therefore $Re(T, S) = \frac{0.44 \times 1}{\sqrt{1+1+1+1}}$, $\epsilon(S) = (0, 0.1, 0.4, -1, 0.1, 0, 0)$. We empirically set the coefficients (ρ, σ, a) to $(0.22, 0.35, 2)$, then $E(S) = 1.47, I(S) = 2$.

4 Experiment

2,302 randomly separated speeches in the dataset are utilized as the training set to estimate the coefficient of each feature. The rest 300 speeches form the testing set. The speeches in training set are annotated as ironical or not, manually through a survey. All interviewees are native speakers of Chinese with good education backgrounds. The theme and speech are presented to three different people for judgement. If more than two people recognized a speech as ironical, it will be annotated as a speech of irony. For the testing set we presented them to five people and annotated it as ironical if three or more people recognized it.

Several statistical machine learning methods can be applied for irony detection. The output space is a binary sequence to determine whether a speech expresses irony or not. Intuitively we choose a logistic regression model to deal with the classification problem. Instead of investigating to each feature respectively as in [12], logistic regression model synthesizes effects of the features. More specifically, we adjust the cut-off point to find out a proper probability threshold to balance between precision and recall. We also compared the model with artificial neural network model and decision tree (C4.5) model. The performance on testing set is illustrated in Figure 1.

The model is implemented with IBM SPSS Statistics 20. $Re(S)$, $E(S)$ and $I(S)$ are used as independent variables. Notably in order to increase the accu-

³ We have referred to the Chinese Dictionary of Emotion, works of the Natural Language Processing and Computational Social Science Lab, Tsinghua University (<http://nlp.csai.tsinghua.edu.cn/site2/index.php/zh/resources/13-v10>).

racy of the model, speech length is a control variable in the model. Descriptive analysis of dependent and independent variables is illustrated as in Table 2 below. As illustrated in the Table, the three features are significantly relevant to the property whether a text is ironic or not. $Re(S)$ and $E(S)$ are negatively correlated. This phenomenon indicates that the deviation of intentional meaning from literal meaning usually occurs with strong emotion.

Table 2. Correlations of Independent and Control Variables

	Mean	Min	Max	S.D.	0	1	2	3
0 Annotation	0.18	0	1	0.306	1.000			
1 Speech Length	28.733	4	191	24.733	-0.85	1.000		
2 $Re(S)$	0.108	0	1.76	0.298	-0.162*	-0.119	1.000	
3 $E(S)$	0.52	0	3	0.824	0.159*	-0.081	-0.328*	1.000
4 $I(S)$	1.79	0	7	5.621	0.138*	0.378	0.042	0.139

4.1 Result and Discussion

Table 2 suggests that all independent variables have a significant impact on the regression. The accuracy of the detection depends on the selection of cut-off point that indicates the significant level. By comparing the outcome of the regression model and the manually annotated result of the testing set, we can draw a confusion matrix to calculate precision, recall and F-measure. Emphasizing on different aspect of precision or recall, F can be written as $\frac{(\beta^2+1)PR}{\beta^2P+R}$. We take parameter $\beta = 0.5$ and $\beta = 1$ into consideration, set the cut-off point from 0.1 to 0.4 with an interval of 0.05. The performance of regression model is listed in Table 3.

Table 3. F-measure with different β and cut-off point

Cut-off	0.4	0.35	0.3	0.25	0.2	0.15	0.1
Recall	9.024%	20.109%	47.890%	61.117%	81.198%	89.402%	95.371%
Precision	47.301%	41.083%	39.771%	37.530%	33.119%	28.467%	27.737%
F ($\beta=0.5$)	25.591%	33.992%	41.166%	40.669%	37.567%	32.960%	32.321%
F ($\beta=1$)	7.578%	13.500%	21.727%	23.251%	23.524%	21.591%	21.487%

From the table we observe that the recall ratio varies violently when different cut-off point is chosen. However, the change of precision is more smooth. Therefore the cost to lift precision is high. When $\beta=0.5$, the F-measure suggests a set of cut-off point to 0.3. When $\beta=1$, the F-measure suggests a set of cut-off point

to 0.2. Under the second circumstance, recall almost doubles, while the decrease of precision is not significant. Thus we set parameter β to 1.

4.2 Evaluation

Table 4. Omnibus test result

	Model 1	Model 2
Speech Length	-0.085*	-0.081**
	[0.07]	[0.009]
<i>Re</i>	—	-1.454***
		[0.311]
<i>I</i>	—	0.272**
		[0.095]
<i>E</i>	—	1.248***
		[0.001]
Constant	-1.671***	-1.650***
	[0.095]	[0.021]
Record Number #	2302	2302
Nagelkerken's R-sqaure	0.186	0.321
Omnibus chi-sqaure	0.149	17.624***

We employed statistical testing to verify the features investigated in our model did impact the efficiency of irony detection. In Omnibus test we compare the two models. Model 1 only includes control variable and constant whereas Model 2 is a model with additional independent variables to consider formalized features. Table 4 illustrates the details⁴. According to Table 4, the three features significantly improve the performance of Model 1.

Figure 1 demonstrates the performance of the model trained in Section 4 on the testing dataset. It is worth pointing out that results in Chinese, English, Italian and Portuguese are not strictly comparable because of the differences in datasets, ratio of ironic texts (in our work 16.0%, in [6] 51.89% and in [5] 12.5%), and testing methods. This may explain the high precision in Portuguese. The Logistic Regression model exceeds ANN and Decision Tree on all the three indicators. Generally speaking, on recall and accuracy our model performed almost as good as in English (the reported accuracy is 71.17% in English, we achieved 60.30% in our model; reported recall ratio in English is 62.17%, and our result outperforms it at 71.21%). If the purpose of detection focuses on computer assisted classification, accuracy should be considered in the first instance. But the precision of irony detection in Chinese is remarkably lower. This phenomenon

⁴ Note: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

suggests that irony detection in Chinese is more challengeable and other forms of expression tend to be misunderstood as irony. In fact, however we adjust the cut-off point, the precision hardly reach 50%.

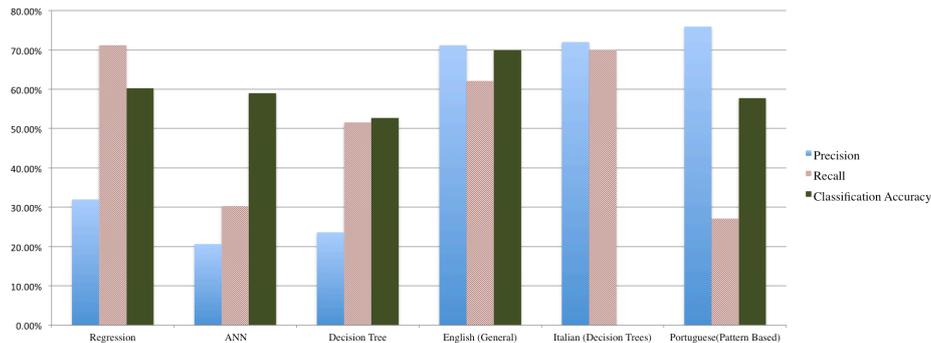


Fig. 1. Precision, Recall and Classification Accuracy of the testing set under different machine learning algorithms. The Logistic Regression, ANN and Decision Tree methods are carried on Chinese dataset. We compare our results with an overall evaluation of several methods in English [12] and results in Italian and Brazilian Portuguese. Accuracy in Italian is not reported.

There are reasons to hamper the irony detection result in Chinese to reach the standard of the other works in English, Italian and Portuguese. One of them could be the low quality of gold standard, which is generated by human annotators. The disagreement on irony is observed more frequently in Chinese during the annotation process. Another reason is the poor performance of natural language processing tools in Chinese. For example, there can be errors in the word segmentation step⁵. A certain word might have different emotional scenarios in different context. Literature [18] enumerates “骄傲 (pride:arrogance)” as an example. This word expresses positive opinion in certain contexts, while negative opinion in other circumstance. This ambiguity has a negative influence on the performance of the emotion dictionary because only the most frequent usage is taken into account.

5 Conclusion

In this paper, we introduce a logistic regression model for irony detection in Chinese online texts. Compared with irony detection in English, this model reinvestigates features of irony from previous research, and computationally formalized them to fit the different language environment. The significance of our features is proved by correlation analysis and Omnibus test. The performance of our model

⁵ Based on this consideration, we have not discussed features involving part-of-speech.

comparing to similar works on English and other European languages suggests that precision of irony detection in Chinese still needs improvement.

Future work includes research on the detection framework and application of this model on analyzing public opinion.

Acknowledgements. This research is partly supported by National High Technology Research and Development Program of China (2012AA011101) and Research Fund of Undergraduate Research and Training Program of Peking University. The authors also want to thank the anonymous reviewers who gave them valuable comments on this work.

References

1. Feng, C: English Rhetorical Options. Foreign Language Teaching and Research Press, Beijing (2011)
2. Utsumi, Akira.: A Unified Theory of Irony and Its Computational Formalization. Volume 1: The 16th International Conference on Computational Linguistics. C 96–2162 (1996)
3. Attardo, S.: Irony as Relevant Inappropriateness. *Journal of Pragmatics* 32, 793–826 (2000)
4. Reyes, A., P. Rosso. Making Objective Decisions from Subjective Data: Detecting Irony in Customer Reviews. *Decision Support Systems* 53, 754–60 (2012a)
5. Francesco Barbieri, Francesco Ronzano, and Horacio Saggion.: Italian Irony Detection in Twitter: a First Approach. *The First Italian Conference on Computational Linguistics CLiC-it 2014*, pp 28–32 (2014)
6. de Freitas, Larissa A., et al.: Pathways for irony detection in tweets. *Proceedings of the 29th Annual ACM Symposium on Applied Computing*. ACM (2014)
7. Burgers, C., M. van Mulken and P. J. Schellens.: Type of Evaluation and Marking of Irony: The Role of Perceived Complexity and Comprehension. *Journal of Pragmatics* 44, 231–42 (2012a)
8. Partington, A.: Phrasal Irony: Its form, Function and Exploitation. *Journal of Pragmatics* 43,1786–800 (2011)
9. Grice, H. Paul.: *Studies in the Way of Words*. Harvard University Press, Cambridge, MA (1989)
10. Giora, Rachel.: On Irony and Negation. *Discourse Processes* 19, 239–264 (1995)
11. Partington, A.: Irony and Reversal of Evaluation. *Journal of Pragmatics* 39,1547–69 (2007)
12. Reyes, A., P. Rosso., T. Veale.: A Multidimensional Approach for Detecting Irony in Twitter. *Language Resources and Evaluation* 47, 239–68 (2013)
13. Reyes, A., P. Rosso.: On the Difficulty of Automatically Detecting Irony: Beyond a Simple Case of Negation. *Knowledge and Information Systems* 40(3), pp 595–614 (2014)
14. Reyes, A., P. Rosso., D. Buscaldi.: From Humour Recognition to Irony Detection: The Figurative Language of Social Media. *Data & Knowledge Engineering* 74, 1–12 (2012b)
15. Wen X.: On the Contextual Cues and Constraints of Ironic Utterances (in Chinese). *Journal of Sichuan International Studies University*. 17(6), 52–55 (2001)

16. Zhu, X.: A Study of Irony from a Pragmatic Point of View (in Chinese). *Journal of Social Science of Hunan Normal University*. 31(3), 128–129 (2002)
17. Wang, M., Shi, H.: Research on sentiment analysis technology and polarity computation of sentiment words. In: 2010 IEEE International Conference on Progress in Informatics and Computing, vol. 1, pp. 331–334. IEEE (2010)
18. Li, R., et al.: A Method of Polarity Computation of Chinese Sentiment Words Based on Gaussian Distribution. *Computational Linguistics and Intelligent Text Processing CICLing 2014, Part II*, pp 53–61 (2014)
19. Yanfen Hao and Tony Veale.: An Ironic Fist in a Velvet Glove: Creative Misrepresentation in the Construction of Ironic Similes. *Minds and Machines* 20(4), 635–650 (2010)
20. Xu, L., Zhang, Q., Wang, D. D. and Zhang, J.: Research of Chinese Segmentation Based on MMSeg and Double Array TRIE. *Advanced Research on Automation, Communication, Architectonics and Materials*, 225-226(1-2), 945–948 (2011)
21. Cruz, F., Troyano, J. A. and Enriquez, F.: Supervised TextRank. *Advances in Natural Language Processing, Proceedings*, 4139, 632–639 (2006)